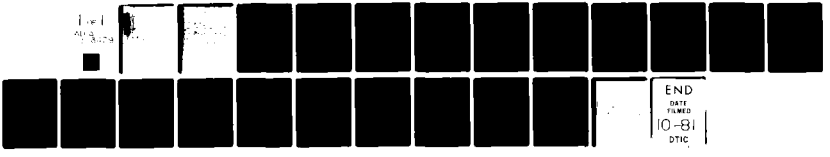
AD-A103 479 CALIFORNIA UNIV BERKELEY CENTER FOR PURE AND APPLIE--ETC F/G 12/1
ANALYTIC COMPLEXITY THEORY AND THE SOLUTION OF AX=B,(U)
JUN 81 B N PARLETT N00014-76-C-0013
UNCLASSIFIED PAM-40 NL



END
DATE FILMED
10-81
DTIC

ANALYTIC COMPLEXITY THEORY AND THE SOLUTION OF $Ax = b$

by B.N. Parlett

Department of Mathematics

and

Computer Science Division

of the EECS Dept.

Date: June 1981

The author is pleased to acknowledge partial support from ONR Contract N00014-76-C-0013. ✓

Abstract. The concepts and results of [1] are discussed in the light of a new result on the optimality of the minimum residual algorithm. The proof is given in Section 3. The terminology of [1] is described in Section 4.

1. Summary

The conjugate gradient algorithm (called CG hereafter) is a popular way to solve large sparse positive definite systems of equations. The minimum residual algorithm (called MR) is closely related to CG and can be applied to any nonsingular system, see [5], [6]. In this summary we describe our result and, of more importance, comment on their significance. The discussion is confined to exact arithmetic because this enquiry concerns only the theory of MR, CG and related algorithms, not their implementation.

Both CG and MR require between 1 and n steps to solve a given n by n nonsingular system $Ax = b$, and both extremes can occur. One of the attractions of these methods is that they are finite and only use A to form products Av. Moreover it is known that both CG and MR are optimal in a certain sense. More precisely, at each step each method produces a vector z which minimizes the norm of the residual vector $b - Az$ over all z in an appropriate subspace of R^n. As usual R^n denotes the space of all real n-dimensional column vectors. The methods differ only in their choice of norm, MR uses the Euclidean norm $\|v\| \equiv \sqrt{v^T v}$ whereas CG uses an "energy" norm $\|A^{-1/2}v\|$. The former is independent of A but the latter has meaning only for symmetric positive definite matrices A. For the sake of simplicity and generality we shall concentrate on MR in the rest of this essay.

Most numerical analysts suppose that the general theory of MR stops at this point. Indeed, to estimate how many steps will be required to

achieve a prescribed reduction in the residual norm some additional information about A, such as its condition number or some other measure of its eigenvalue distribution, must be supplied.

Recently, in [1], Traub and Wozniakowski have sought to enlarge the scope of the discussion of MR. Following the lines laid down in [2] they remove the restriction of algorithms to the standard (polynomial) class, over which MR and CG are optimal. Why, they say, restrict attention to the standard (Krylov[†]) subspace of R^n ?

Will MR remain optimal when any rival algorithm is allowed? They address this question in [1] while erecting a new framework around the study, in general, of iterative algorithms for solving $Ax = b$. They succeed in showing that MR costs at most one iteration-step more than an optimum algorithm could require.

Actually MR is still optimal for positive definite A; this note will explain why, and then go on to indicate how all such results, theirs and ours, can mislead the unwary reader.

2. Discussion of Ref. [1]

To go any further we must introduce some terminology. At step j MR computes a linear combination x_j of the vectors b, Ab, A^2b, ..., $A^{j-1}b$. No generality is lost by the assumption (in force throughout the paper) that this set of vectors is linearly independent.

[†] Krylov subspaces are defined in Section 2.

Another vector $A^j b$, or something equivalent to it, is needed to determine the coefficients for MR's output x_j although $A^j b$ itself is not part of the linear combination. For future reference we let ν_j denote x_j's residual norm. The defining property of MR and an expression for x_j are given in the following equations,

$$\begin{aligned} \nu_j &\equiv \min_{\gamma_i} \|b - \gamma_0 Ab - \gamma_1 A^2 b - \ldots - \gamma_{j-1} A^j b\| , \\ &= \min_{\gamma_i} \|b - A(\gamma_0 b + \gamma_1 Ab + \ldots + \gamma_{j-1} A^{j-1} b)\| , \\ &= \|b - Ax_j\| . \end{aligned}$$

The theory in [1] can be motivated by asking whether there exists some clever, well-hidden algorithm which, with no further multiplication of vectors by A , delivers a vector z with $\|b - Az\| < \nu_j$ and so surpasses MR . The answer, after a little thought, is yes. The "well hidden" algorithm is Gaussian elimination (with pivoting) which delivers $A^{-1}b$ itself without computing Av for any v ! Of course we feel cheated and this reaction shows that the question must be posed with more care. Rival algorithms must be prevented from getting at A through the back door.

One way to be fair is to limit knowledge of A to the current information: b , $Ab, \ldots, A^j b$, or some "equivalent" set of vectors. In general these vectors do not fix A uniquely and so there is the set

$$A^j = \{\tilde{A}: \tilde{A}^i b = A^i b , \quad i = 1, \ldots, j\}$$

of matrices indistinguishable from A with the given information. Sometimes we restrict A^j to a particular class of matrices $\tilde{A}$, sometimes not. The context decides.

Traub and Wozniakowski use A^j in specifying the cost of an algorithm for a given matrix. The technical description is given in Section 4 (along with a warning) but, in English, the cost is the minimal number of steps j required to achieve a given reduction in the residual norm for any $\tilde{A}$ in A^j, i.e. the number of steps needed in the worst case. This cost, for an algorithm ϕ, is denoted by $k(\phi, A)$. The new version of optimality on which [1] is based is optimality over a whole class F of matrices. The authors seek an algorithm $\bar{\phi}$, which satisfies

$$k(\bar{\phi}, A) = \min_{\phi} k(\phi, A) \quad \text{for each } A \text{ in } F .$$

The minimum is over all algorithms using the given information. These are the basic concepts. There are many theorems in [1]. The main result, as mentioned above, is that over the class $F = SPD$, the class of symmetric positive definite matrices, and also over larger classes F, MR is within one matrix-vector multiplication of being an optimal algorithm; but [1] does not exhibit an algorithm $\bar{\phi}$ that is optimal for any of those classes F. Instead they study restricted classes and finish [1] with conjectures and open problems, as if the whole field were ripe for further development.

We doubt the ripeness. The questions being asked seem to be the wrong questions. For a start the quest for many optimal $\bar{\phi}$ is not necessary because

MR is optimal for the class SPD ,

and for any larger class.

The proof is given in Section 3. It amounts to showing that for each rival vector $y \neq x_j$ one can construct a positive definite $\bar{A} = \bar{A}(y)$ in A^j such that

$$\|b - \bar{A}y\| > \nu_j .$$

Thus y is worse than x_j for $\bar{A}$ and for many other matrices in A^j . Even worse, for y not in the usual subspaces, $\{\|b - \tilde{A}y\| : \tilde{A} \in A^j \cap \text{SPD}\}$ is unbounded.

The devastating consequence of this observation for the theory launched by Traub and Wozniakowski is that if ϕ produces any approximation unrestricted to combinations of b , Ab, ..., $A^{j-1}b$ then for A's in SPD the cost $k(\phi, A) \geq n$.

In other words MR wins because all its new rivals die at the starting gate. The villain of the piece is the class SPD . It is so big that the generality, espoused in [1], of allowing any algorithm ϕ (and hence approximations outside the usual subspace) is annulled by the required fairness of taking the worst case in A^j . The theory becomes vacuous.

On the other hand SPD is the most important class of matrices in the application of iterative methods.

One way to try to save the theory is to restrict it to small subsets of SPD , for example to those matrices with condition number ≤ 100 . Yet even here $k(\phi, A) = n$ unless ϕ is very close to MR , as shown in Section 3.

In those cases when $k(\phi, A) < n$ the theory in [1], though not vacuous, reduces to a translation into its terminology of standard bounds, see [4] for example, on convergence rates of MR . How sad it is to evade the Charybdis of vacuity only to founder on the Scylla of redundancy.†

What has gone wrong? Why has such an enticing research program, on closer examination, dissolved away like the Cheshire Cat in "Alice Through the Looking Glass"? We are not sure but we offer the following thought. The information $\{b , Ab , A^2b, \ldots, A^jb\}$ seems to arise as an artifact of the MR and CG algorithms, not as part of the <u>problem</u> of solving $Ax = b$ by some iterative method. The optimality of MR over SPD may be true but it is no more exciting than the discovery that the dress suit made for Mr. X fits him better than it does anyone else.

For readers of this note who are not familiar with [1] we have added a final section in which its leading notions are stated, together with our comments. This will make it clear that this paper and [1] are concerned with exactly the same problem, namely the analytic complexity of methods for solving $Ax = b$ which are restricted to the use of Krylov information on A . Whereas [1] claims that this is a new topic deserving of development we suggest that the nonvacuous aspects are well known, although they are usually expressed in more mundane language than is favored in [1].

Here, in one sentence, is our version of the analytic complexity of the problem. Over the standard matrix classes (SPD, Sym, nonsingular), among all algorithms using the given information MR is (worst case) optimal and its complexity (number of steps required) is n , which is as bad as it could be.

---

†From the Greek myths. Boats in the Strait of Messina had to steer a very fine line between two monsters Charybdis (a whirlpool) and Scylla (a rock).

Unless worst case analyses yield something better than worst possible outcomes our interest shifts from them to the reasons why particular cases fare so well. Except for special right hand sides b the residual reduction produced by MR depends entirely on the usually unknown distribution of eigenvalues. Significant diminutions occur after a small number of steps when the eigenvalues are grouped into a small number of clusters — even when the condition number is large. It is specific surprising results such as this which give substance to the study of MR and CG .

3. Optimality of MR

A little preparation is necessary before we can state and prove the theorem.

The only information available concerning the nonsingular n by n matrix A is the set of vectors $\{b\ ,\ Ab\ ,\ A^2b, \ldots, A^jb\}$, $1 \leq j < n$. Actually the methods use "equivalent" combinations of these vectors which yield alternative, and more practical bases for the subspace

$$K^{j+1} = \text{span}\{b\ ,\ Ab, \ldots, A^jb\}\ ,\quad 1 \leq j < n\ .$$

These subspaces of $\mathbb{R}^n$ are sometimes called Krylov subspaces.

In theory it is possible to have A^mb dependent on b , Ab, ..., A^{m-1} for some $m < n$. In this case it is easily verified that K^m is invariant under A . Moreover the exact solution $A^{-1}b$ lies in K^m . Consequently

there is no need to look outside K^m for an approximation. Theoretically A can be replaced by its restriction to K^m and the change will not be noticed. So, without loss of generality, we may invoke

Assumption 1. $\{b, Ab, \ldots, A^{n-1}b\}$ is linearly independent.

Now the superscripts on the K^j indicate dimension.

A useful basis for K^n ($= \mathbb{R}^n$) is the Lanczos basis $\{q_1, q_2, \ldots, q_n\}$ obtained by applying the Gram-Schmidt orthonormalizing process to the original ordered basis $\{b, Ab, \ldots, A^{n-1}b\}$. If we introduce the n by j matrix $Q_j = (q_1, q_2, \ldots, q_j)$ we have

$$Q_j^T Q_j = I_j \text{ , the } j \text{ by } j \text{ identity matrix, } j = 1, \ldots, n . \qquad (2\text{-}1)$$

By representing A in this basis our proof becomes transparent.

The actual details of the MR algorithm are not important here. Its output x_j, at step j, depends on $A^j b$ but x_j itself lies in K^j. By definition of MR x_j satisfies

$$\nu_j \equiv \min_{v \in K^j} \|b - Av\| = \|b - Ax_j\| . \qquad (2\text{-}2)$$

Note that the information available determines the action of A on K^j but not on K^{j+1} since $A \cdot A^j b$ is unknown.

The set of matrices indistinguishable from A at step j is

$$A^j = \{\tilde{A}: \tilde{A}^i b = A^i b , \ i = 1, \ldots, j\} , \ j < n .$$

In other words, at this step K^{j+1} is uniquely determined but K^{j+2} is not (unless $j = n - 1$).

It will be convenient to abbreviate by GL, Sym, and SPD the classes of nonsingular, symmetric, and symmetric positive definite matrices.

The reason for introducing the Lanczos basis is that with respect to it any $A \in \text{Sym}$ is represented by a tridiagonal matrix T_n. We define

$$T_j = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_j \\ & & & \beta_j & \alpha_j \end{bmatrix} \tag{2-3}$$

The preparation is over. We now proceed to the statment and proof of the theorem.

The optimality of MR for a single matrix A has been stated above in (2-2), but MR is optimal in a somewhat broader sense. With repect to $A^j \cap \text{SPD}$ there is no vector in $\mathbb{R}^n$ which is better than x_j because of the

THEOREM. $$\min_{y \in \mathbb{R}^n} \sup_{\tilde{A} \in A^j \cap \text{SPD}} \|b - \tilde{A}y\| = \nu_j .$$

We prove the theorem by considering a special one-parameter family of matrices A_α and establishing two lemmas. Using the orthonormal basis consisting of the columns of Q_n we have

$$Q_n^T A Q_n = \left[\begin{array}{c|c} T_j & 0 \\ \hline 0 & U \end{array}\right] \qquad (2\text{-}4)$$

where $\cdot = \beta_{j+1}$ and is known, whereas U stands for the unknown part of T_n. Different $\tilde{A} \in A^j \cap \text{SPD}$ will have different values for U. We define A_α implicitly by

$$Q_n^T A_\alpha Q_n = \left[\begin{array}{c|c} T_j & 0 \\ \hline 0 & \alpha I_{n-j} \end{array}\right] \qquad (2\text{-}5)$$

where $\cdot = \beta_{j+1}$.

LEMMA 1. For all large enough α, $A_\alpha \in A^j \cap \text{SPD}$.

Proof. Recall that Q_n is orthogonal so that $A_\alpha \in \text{Sym}$, by construction. Moreover $A_\alpha \in A^j$ by comparison of (2-4) and (2-5). More formally, writing e_i for the i^{th} column of I, we have

$$A_\alpha q_i = Q_n T_j e_i + \delta_{ij}\beta_{j+1} q_{j+1} = A q_i \,, \quad i \leq j \,.$$

The Kronecker symbol δ_{ij} is the i^{th} element of e_j. Thus

$$A_\alpha K^i = A K^i \,, \quad i \leq j \,,$$

and so A_α and A produce the same K^i for $i = 1, \ldots, j+1$.

It remains to see when $A_\alpha \in SPD$. Note that by doing a block triangular factorization

$$Q_n^T A_\alpha Q_n = \left[\begin{array}{c|c} I_j & 0 \\ \hline c^T & \\ 0 & I_{n-j} \end{array}\right] \left[\begin{array}{c|c} T_j & 0 \\ \hline 0 & \alpha I_{n-j} - \beta_{j+1}^2 \eta_j e_1 e_1^T \end{array}\right] \left[\begin{array}{c|c} I_j & c\ 0 \\ \hline 0 & I_{n-j} \end{array}\right]$$

where $c^T = \beta_{j+1} e_j^T T_j^{-1}$ and $\eta_j = e_j^T T_j^{-1} e_j$. In other words A_α is congruent to the direct sum

$$T_j \oplus \mathrm{diag}(\alpha - \beta_{j+1}^2 \eta_j \ , \ \alpha, \ \ldots, \ \alpha) \ .$$

Observe that

$$0 < \eta_j \leq \|T_j^{-1}\| = 1/\lambda_{\min}(T_j) \leq 1/\lambda_{\min}(A) \ .$$

So $A_\alpha \in SPD$ when $\alpha > \beta_{j+1}^2/\lambda_{\min}(A)$. □

LEMMA 2. If $y \notin K^j$ then $\|b - A_\alpha y\| \to \infty$ as $\alpha \to \infty$. For each $y \in R^n$ there is an $\bar{\alpha}$ such that $\|b - A_\alpha y\| \geq \nu_j$ for $\alpha \geq \bar{\alpha}$, and equality holds if an only if $y = x_j$.

Proof. Represent y in the Lanczos basis as

$$y = Q_n \begin{pmatrix} s \\ t \end{pmatrix}$$

where $s \in R^j$, $t \in R^{n-j}$. Then, with $b = q_1\beta_1$,

$$\|b - A_\alpha y\| = \left\| Q_n e_1\beta_1 - Q_n \left[\begin{array}{c|c} T_j & \\ \hline & \alpha I_{n-j} \end{array}\right] \begin{pmatrix} s \\ t \end{pmatrix} \right\| ,$$

$$= \left\| \begin{pmatrix} e_1\beta_1 - T_j s - e_j\tau \\ - e_1\sigma - \alpha t \end{pmatrix} \right\| , \quad \text{since } Q_n \text{ is orthogonal,}$$

where $\tau = e_1^T t$, $\sigma = e_j^T s$.
If $y \notin K^j$ then $t \neq 0$ and

$$\|b - A_\alpha y\| \geqslant \alpha \|t + e_1\sigma/\alpha\| ,$$
$$\to \infty , \quad \text{as} \quad \alpha \to \infty .$$

More precisely for any $y \notin K^j$, $\|b - A_\alpha y\|^2$ is a quadratic in α with leading coefficient $\|t\|^2$ and so exceeds ν_j^2 for all α exceeding some $\bar{\alpha}$ depending on t and s . When $t = 0$ then $A_\alpha y = Ay$, we can take $\bar{\alpha} = 0$, and standard least squares theory states that the minimum residual norm ν_j is attained when $b - Ay \perp K^j$. This occurs only when $s = T_j^{-1} e_1\beta_1$. In that case

$$y = x_j = Q_j s \quad \text{and} \quad \nu_j = |e_j^T T_j^{-1} e_1|\beta_1 . \quad \square$$

Proof of theorem.

If $y \notin K^j$ then by Lemmas 1 and 2,

$$\sup_{A^j \cap SPD} \|b - \tilde{A}y\| \geq \sup_{\alpha} \|b - A_\alpha y\| = \infty .$$

If $y \in K^j$ then

$$\sup_{A^j \cap SPD} \|b - \tilde{A}y\| = \|b - Ay\| \geq \|b - Ax_j\| = \nu_j . \quad \square$$

We remark that we have actually proved that MR is optimal over any class larger than SPD, in particular over Sym and GL.

Let us consider briefly what happens when α is not permitted to go to ∞. Using the expressions in the proof of Lemma 2 we write $u = e_1\beta_1 - T_j s$ and find that

$$\|b - A_\alpha y\|^2 = \|u\|^2 - 2\mu\tau + \tau^2 + \sigma^2 + 2\alpha\sigma\tau + \alpha^2\|t\|^2$$

where $\mu = e_j^T u$, $\tau = e_1^T t$, $\sigma = e_j^T s$. It would be unreasonable to demand $\alpha < \|A\|$ and so we see that, unless $\|t\|$ is tiny, i.e., ϕ is close to MR, we can have $\|b - A_\alpha y\| \geq \|A\| \cdot \|t\| >$ the required accuracy. In such cases we still have $k(\phi, A) = n$ and the theory in [1] becomes vacuous.

4. The Complexity Connection

We describe the principal terms in [1].

The task is to find x in $\mathbb{R}^n$ so that, for given A and b, $\|b - Ax\| < \varepsilon\|b\|$, for some given fixed ε. Thus ε is a parameter in the theory. The information $\{b, Ab, \ldots, A^j b\}$ is denoted by $N_j(A, b)$.

The class from which A is drawn is F. Our set

$$A^j = \{\tilde{A}: \tilde{A} \in F , \ \tilde{A}^i b = A^i b , \ i = 1, \ldots, j\}$$

is denoted by $V(N_j(A, b))$. Recall, from [3] or [4], that MR is optimal among those algorithms which use N_j and deliver output in the Krylov subspace K^j. The springboard for [1] is the removal of this restriction on rival algorithms. We shall say that an algorithm is <u>new</u> if its output using N_j is not in K^j. Now we are ready to consider the main concepts in [1].

For any algorithm ϕ, whose output from N_j is z_j, [1] defines

$$k(\phi, A) \equiv \min\{k: \|b - Az_k\| < \varepsilon\|b\| , \ \forall \ \tilde{A} \in A^k\}$$

as the <u>matrix index</u>. It purports to be the cost of ϕ as measured by the number of matrix-vector products needed in the worst case. Unfortunately an enlargement in F can increase $k(\phi, A)$ thus robbing $k(\phi, A)$ of that meaning. As shown in Section 2 we may assume that the set N_j is linearly independent for all $j < n$. When $j = n$ the set must be dependent (it has $n + 1$ vectors) and MR delivers the exact solution. Our theorem in Section 2 shows that if $F \supset SPD$ and ϕ is a new algorithm then $\sup_{\tilde{A} \in A^k} \|b - \tilde{A}z_k\| = \infty$, for all $k < n$. Hence $k(\phi, A) \geq n$!

For old algorithms ψ $k(\psi, A) \geq k(MR, A)$. The value of ε is irrelevant.

Next comes the class index of ϕ ,

$$k(\phi, F) = \sup_{A \in F} k(\phi, A) .$$

For new algorithms $k(\phi, F) \geq n$ for $F \supset SPD$ and for many smaller classes. Even for MR , $k(MR, SPD) = n$. Since $k(\phi, A)$ itself depends on F the distinction between $k(\phi, A)$ and $k(\phi, F)$ is obscure.

Traub and Wozniakowski seek to minimize these indices. They define the optimal matrix index by

$$k(A) = \min_{\phi} k(\phi, A)$$

and the optimal class index by

$$k(F) = \max_{A \in F} k(A) .$$

By the remarks made above, if $F \supset SPD$ then

$$\min_{\text{new } \phi} k(\phi, A) \geq n$$

while

$$\min_{\text{old } \psi} k(\psi, A) = k(MR, A) .$$

An algorithm $\bar{\phi}$ is called strongly optimal (over F) if

$$k(\bar{\phi}, A) = k(A) , \quad \forall A \in F ,$$

and is called optimal (over F) if

$$k(\bar{\phi}, F) = k(F) .$$

We have just seen that MR is strongly optimal over any $F \supset SPD$.

It should be mentioned that [1] does not exhibit this result. Instead a nice proof shows that if F is orthogonally invariant then

$$k(MR, A) - k(A) \leqslant 1 , \quad \forall\ A \in F .$$

In addition a very contrived 2 by 2 example is given to show that equality can occur for a peculiar F.

We want to make a comment on the hypothesis that F be orthogonally invariant. One of the attractive features of Krylov information $N_j(A, b)$ is that it is coordinate free. The standard methods and theory are geometric in nature. Any change of basis which preserves the Euclidean norm is permitted. It seems unnatural to introduce matrix classes F which do not share this property with the information. It is hardly surprising that the study made in [1] of tridiagonal matrices shows that "anything can happen."

By studying subclasses of SPD whose members have condition numbers bounded by small enough κ it is possible to obtain $k(\phi, A) < n$, $\forall A \in F$ and sometimes $k(A) < n$. Of course $k(MR, A)$ will then be a complicated function of ε and κ. This topic has been studied by numerical analysits. They showed how $cond(A)$ can be used to bound $k(MR, A)$. With more information on the spectrum more accurate bounds can be derived. Unfortunately the theorems in Section 5 on symmetric matrices contain no

new insights but are translations into the language of complexity of the known bounds mentioned above. In contrast the results on unsymmetric matrices are new and quite misleading. The authors prove in a dead pan manner that even when $cond(A) = 1$ it is possible to have $k(MR, A) = n$. Not a word is said about the fact that for unsymmetric matrices $cond(A)$ is no indicator of the clustering of A's spectrum and so is virtually irrelevant to the convergence rate of MR. $Cond(A)$ measures the spread of singular values, not the eigenvalues.

The theorems in [1] tell us more about Analytic Complexity theory than about iterative methods for solving $Ax = b$.

Acknowledgement

I would like to thank Jim Demmel and Velvel Kahan, who have suggested improvements in the presentation of the ideas in this note.

References

[1] J.F. Traub and H. Wozniakowski, "On the optimal solution of large linear systems", Technical Report, Comp. Sci. Dept., Columbia University, New York, January 1980.

[2] _____, A General Theory of Optimal Algorithms, ACM Monograph series, Academic Press, 1980.

[3] M. Hestenes and E. Stiefel, "Methods of Conjugate Gradients for solving linear systems," J. Res. Nat. Bur. Standards 49(1952) 409-436.

[4] S. Kaniel, "Estimates for some computational techniques in linear algebra", Math. Comp. 20(1966), 369-378.

[5] C.C. Paige and M.A. Saunders, "Solution of sparse indefinite systems of linear equations," SIAM Jour. Num. Anal. 12(1975) 617-629.

[6] C. Lanczos, "Solution of Systems of Linear Equations by Minimized Iterations," J. Res. Nat. Bur. Standards 49(1952) 33-53.